A NEW SEMANTIC SEGMENTATION MODEL FOR REMOTE SENSING IMAGES

Xin Wei[1]*, *Yajing Guo*[2], *Xin Gao*[1], *Menglong Yan*[1], *Xian Sun*[1]

[1]Key Laboratory of Technology in Geo-spatial Information Processing and Application
System, Institute of Electronics, Chinese Academy of Sciences
[2]Beijing University of Posts and Telecommunications
{weixin215@mails.ucas.ac.cn}

ABSTRACT

Semantic segmentation for remote sensing images is a critical process in the workflow of object-based image analysis. Recently, convolutional neural networks(CNNs) are powerful visual models that yield hierarchies of features. In this paper, we propose a deep convolutional encoder-decoder model for remote sensing images segmentation. Specifically, we rely on the encoder network to extract the high-level semantic feature of ultra-high resolution images and the decoder network is employed to map the low resolution encoder feature maps to full input resolution feature maps for pixel-wise labeling. Also the fully connected conditional random field (CRF) is integrated into the model so that the network can be trained end-to-end. Experiments on the Vaihingen dataset demonstrate that our model can make promising performance.

Index Terms— Convolutional neural networks, conditional random field, semantic segmentation, remote sensing images

1. INTRODUCTION

Recently, object-based image analysis for high resolution remote sensing images has become a crucial and evolving topic in photogrammetry. During the workflow of image analysis, semantic segmentation is a prerequisite procedure for subsequent object recognition and information extraction [1]. The traditional methods of image segmentation can be classified into edge-based [2], region-based [3] and hybrid segmentation. However, since convolutional neural networks (CNNs) have shown excellent performance in numerous visual recognition tasks such as object classification, object detection and semantic segmentation, the CNNs-based semantic segmentation has become increasingly popular.

Semantic segmentation, as one of the classical computer vision tasks, aims to solve structured pixel-wise labeling problems. The state-of-the-art algorithms convert an existing CNN architecture designed for object classification to a fully convolutional network (FCN) [4]. A pool of convolutional layers are employed to extract the high-level semantic features of input images, so they can only obtain a coarse feature map with low resolution. There are a couple of critical limitations in FCN-based segmentation. First, due to the fixed receptive field the fully convolutional networks can only be used to solve the single scale objects. Second, the results are too coarse with a simple deconvolution procedure. To overcome such limitations, Noh et al. [5] proposed a multi-layer deconvolution network, which is composed of deconvolution, unpooling, and rectified linear unit (ReLU) layers. However, to obtain instance-wise segmentations, the training process is complicated. Conditional random field (CRF) [6] and markov random field (MRF) which are post-processing techniques for semantic segmentation can boost the performance.

To apply semantic segmentation into remote sensing images, we proposed a new architecture which take advantage of the symmetry mechanism of deconvolution network and the simplicity of fully convolutional network. In addition, we integrate the fully connected conditional random field into our network and this is our main contribution. In such way, the whole segmentation network can be trained end-to-end. We evaluate our model on a remote sensing images dataset called Vaihingen with promising results.

The remainder of the paper is organized as follows. In Section 2, we elaborate our method and network architecture. We analyze our model and evaluate the performance on a remote sensing images dataset in Section 3. This is followed by a general conclusion in Section 4.

2. METHOD

In this section we first briefly describe the architecture of segmentation model. Then, we elaborate the conditional random field algorithm in order to improve the performance.

2.1. Deep Convolutional Encoder-Decoder Architecture

Our model mainly consists of two modules. The first module is a deep convolutional encoder-decoder architecture, and

*The study described in this paper is funded by the National Natural Science Foundation of China(Grant No.41301493).

978-1-5090-4951-6/17/$31.00 ©2017 IEEE

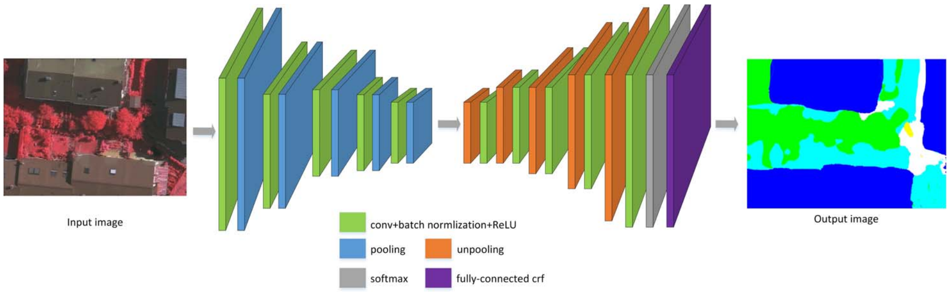


Fig. 1. **Our semantic segmentation architecture.** The layers in green is convolutional layers indicated as "conv". The VGG-16 networks [7] have 5 group convolutional layers (each group contains 2 or 3 convolutional layers). The encoder module and decoder module are completely symmetrical. The layers in purple are fully connected crfs which take the segmentation map generated by softmax layer as input and output refined segmentation. **Best viewed in color**.

the second module takes the output of first module as input and outputs refined results. The whole network is shown as Fig.1. For the deep convolutional encoder-decoder network, the encoder module is based on VGG-16 networks [7]. We remove all the fully connected layers and insert batch normalization layer between each convolutional layer and relu layer. Different from the VGG-16 networks which do not include unpooling layers, the max pooling layer of our model will record the indices of maximum activations selected during pooling operation. In the corresponding decoder module, the unpooling layer puts each activation back to its original pooled location and pads zero in the other location. This strategy can effectively restore the structure of input object. Each unpooling layer is followed by a set of convolutional layers whose number is equal to that of corresponding layers in the encoder module. The decoder module and encoder module are totally symmetrical. Following the decoder module, we use a softmax layer to classify every pixel, and the output is a C channel segmentation map where C is the number of classes.

2.2. Conditional Random Field

The result of the encoder-decoder module is merely a coarse pixel-wise image labelling in which numerous pixels are misclassification especially around the boundary. To improve the result, we integrate a fully connected pairwise conditional random field (CRF) layer into our model. The relationship between image I(global observation) and it's corresponding labels X can be modeled as a Gibbs distribution,

$$P(X = x|I) = \frac{1}{Z(I)} e^{-E(x|I)} \tag{1}$$

For convenience, we will omit the condition I in the notation hereinafter. The energy of a label assignment x is given by:

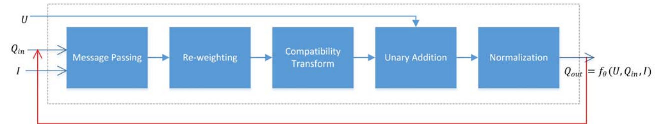


Fig. 2. **An iteration for CRF as a stack of common CNN layers.** U is the pixel-wise unary potential value output by former softmax layer and Q_{in} is the estimation of marginal probabilities from previous iteration.

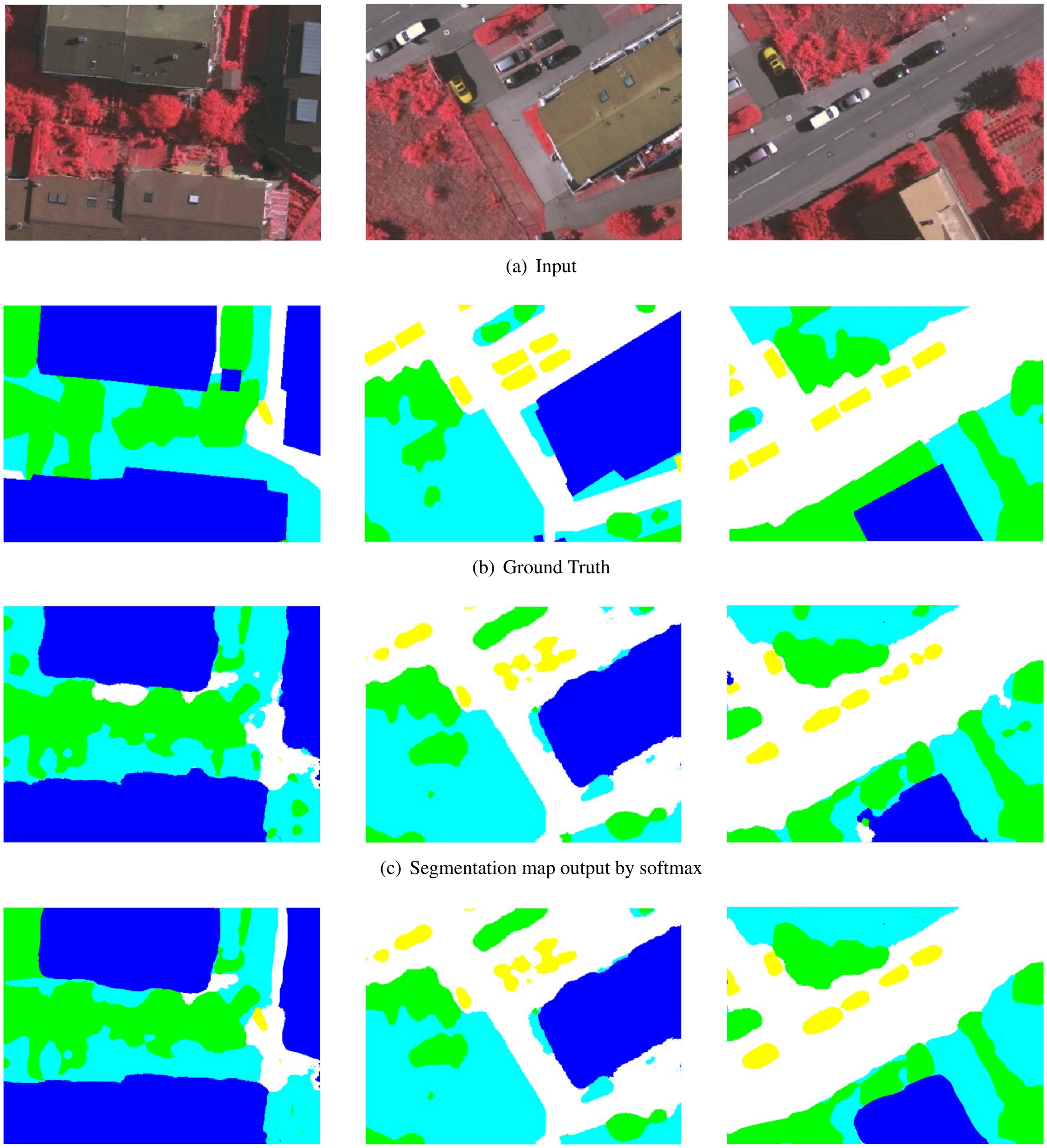
$$E(x) = \sum_{i} \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{2}$$

where the $\psi_u(x_i)$ denotes the inverse likelihood of the pixel i taking the label x_i, and the $\psi_p(x_i, x_j)$ denotes the cost of assigning label x_i, $x_j$to pixel i, j simultaneously. We use the softmax layer's output as the $\psi_u(x_i)$ value. The $\psi_p(x_i, x_j)$ term demonstrates that pixels with similar properties(such as location and color) are assigned with the same labels with high probability, so we formulate it by weighted Gaussians:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} \omega^{(m)} K_G^{(m)}(f_i, f_j) \tag{3}$$

where f_i is feature vector of pixel i, and we use spatial position and RGB color adopted in [6] as the similar properties. $K_G^{(m)}, (m = 1, 2, 3)$ is Gaussian kernel of feature vector f, and the $\mu(x_i, x_j)$ captures the compatibility between different pair labels.

By minimizing the energy $E(x)$, the label with highest probability can be calculated by Equation 1. To minimize $E(x)$, we approximate the CRF distribution $P(X)$ with a simpler distribution $Q(X)$.



(a) Input

(b) Ground Truth

(c) Segmentation map output by softmax

(d) Refined map by fully connected CRF

Fig. 3. Experimental Results. (a) Input images. (b) Ground Truth. (c) Segmentation map output by softmax. (d)Refined map by fully connected CRF. **Best viewed in color**.

To integrate CRF into our model, we use the recurrent neural networks architecture [8]. The general pipeline is shown in Fig.2

3. EXPERIMENTS

3.1. Vaihingen Dataset

Our experiments are based on “semantic labeling contest” of ISPRS WG III/4 Vaihingen dataset. Vaihingen is a relatively small village with many detached buildings and small multi story buildings. The dataset contains 33 patches (of different sizes), each consisting of a true orthophoto (TOP) extracted from a larger TOP mosaic. The ground sampling distance of

Table 1. Segmentation performance on Vaihingen dataset. All models are trained on the same training set. F1-score is adopted to evaluate the performance.

model	Average F1	building	tree	clutter/background	low vegetation	car	impervious surfaces
model without CRF	77.51	87.80	84.5	76.5	71.82	75.33	82.11
model with separated CRF	78.82	89.22	87.77	77.12	73.28	75.42	83.34
our model	**81.77**	**91.22**	**88.22**	**77.98**	**75.48**	**77.73**	**84.59**

TOP is 9 cm and saved as 8 bit TIFF files with three bands (near infrared, red and green bands).

We split the 16 labeled images into two set, including 11 images for training and the others for testing. The original image is too large whose resolution is more than 9000 × 9000, therefore we randomly split each image into 30 subimages with the size of 480 × 360. These subimages are also randomly flipped and rotated to augment the training data. Eventually, we obtain 660 images for training and 300 images for testing.

3.2. Experimental Results

We conduct a series of three experiments to validate the performance of our model. The first set of experiments is based on only deep convolutional encoder-decoder networks. And the second experiment is finetuning the results of the first experiment by a separated CRF. The third set of experiments adopts our ultimate model which integrate the CRF into network as a recurrent neural network. The softmax loss function is employed for pixel-wise classification loss. The training process uses stochastic gradient descent (SGD) to minimize the loss for tuning the model. We set the momentum parameter to be 0.99 and set the initial learning rate at 0.001. We run SGD for 40k mini-batch iterations in all, and the results are shown in Table 1. The results demonstrate that integrating CRF into the network can effectively enhance segmentation performance.

In Table 2, we report the segmentation performance of our proposed model comparing with other methods. Experiments on the Vaihingen dataset demonstrate that our model can make promising performance.

Table 2. Comparison with other methods. We adopt average accuracy (AA) and average F1-score to evaluate the performance.

method	average F1	AA
superpixels-based	69.25	67.15
object proposal-based	76.82	73.76
ours	**81.77**	**76.32**

4. CONCLUSIONS

In this paper, we propose a new semantic segmentation model for remote sensing images. The approach mainly relies on two key factors: 1) deep convolutional encoder-decoder architecture can efficiently extract the high-level semantic feature of the input images and map low resolution feature maps to full input resolution, 2) conditional random field (CRF) can assign same labels to the pixels that share similar properties. We merge the encoder-decoder networks and fully connected CRF into a whole. Experiments on Vaihingen dataset show good performances of the proposed model.

5. REFERENCES

[1] J. Yang, Y. He, J. Caspersen, and T. Jones, “A discrepancy measure for segmentation evaluation from the perspective of object recognition,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 101, pp. 186–192, 2015.

[2] D. Li, G. Zhang, Z. Wu, and L. Yi, “An edge embedded marker-based watershed algorithm for high spatial resolution remote sensing image segmentation,” *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2781–2787, 2010.

[3] X. Zhang, P. Xiao, X. Feng, J. Wang, and Z. Wang, “Hybrid region merging method for segmentation of high-resolution remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, pp. 19–28, 2014.

[4] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE CVPR*, 2015, pp. 3431–3440.

[5] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *IEEE ICCV*, 2015, pp. 1520–1528.

[6] V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” *Adv. Neural Inf. Process. Syst*, 2011.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[8] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P.HS Torr, “Conditional random fields as recurrent neural networks,” in *IEEE ICCV*, 2015, pp. 1529–1537.